DOCUMENT RESUME

ED 453 260 TM 032 797

AUTHOR Elmore, Patricia B.; Rotou, Ourania
TITLE A Primer on Basic Effect Size Concepts.
PUB DATE 2001-04-14
NOTE 13p.; Paper produced at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Information Analyses (070) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; *Research Design; *Statistical Significance

ABSTRACT

The increased interest in reporting effect sizes means that it is necessary to consider what should be included in a primer on effect sizes. A review of papers on effect sizes and commonly repeated statistical analyses suggests that it is important to discuss effect sizes relative to bivariate correlation, t-tests, analysis of variance/covariance, and multiple regression/correlation. An agreed upon nomenclature regarding effect sizes should be established. R. Rosenthal (1994) has classified effect sizes into the "r" family (the Pearson product moment correlation coefficient and the various squared indices of "r" and "r"-type quantities) and the "d" family (mean difference and standardized mean difference indices). Other measures of effect size have been suggested, and some suggestions are given for further reading on these measures. Parsimony and replication should be joined by meaning as principles to consider in reporting research results. To enhance meaning and interpretability of research findings, it is essential that various psychometric variables and test scores be studied and reported for specific samples under varied conditions. (Contains 29 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 260

A Primer on Basic Effect Size Concepts

Patricia B. Elmore and Ourania Rotou
Southern Illinois University Carbondale

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
P. B. Elmore
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM032797

Paper Presented at the Annual Meeting of the
American Educational Research Association
Seattle

April, 2001

BEST COPY AVAILABLE

A Primer on Basic Effect Size Concepts

The issue of required reporting of effect sizes when statistical significance tests are reported in journal articles has been a debate in scholarly publications and at annual meetings of both the American Educational Research Association (AERA) and the American Psychological Association (APA). This symposium entitled "Effect Size Indices: How We Know (How Much) We Know" sponsored by Division D, Measurement and Research Methodology, is certainly appropriate given the theme of the 2001 AERA Annual Meeting, "What We Know and How We Know It," selected by AERA President, Catherine Snow.

Any discussion of this topic should begin with the publication of the report of Leland Wilkinson and the Task Force on Statistical Inference in American Psychologist (1999) that provides guidelines and explanations for the application and reporting of statistical methods in psychology journals. Our comments today are based on two scientific premises central to that report and to our work as scientists, parsimony and replication. Concerning parsimony, they state: "If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as to theories" (p. 598). With regard to replication and stability of findings, they state: "We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. Reporting effect sizes also informs power analyses and meta-analyses needed in future research" (p. 599). The last sentence of the report provides a framework for considering the comments and presentations that are a part of this symposium: "Statistical

methods should guide and discipline our thinking but should not determine it" (p. 603).

Is the idea of reporting effect sizes new or has it been an accepted procedure by statisticians for many years? In the seventh edition of Design of Experiments Fisher (1960) stated that "convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1% we do not ... ever need to lose sight of the exact strength which the evidence has in fact reached, or to ignore the fact that with further trial it might come to be stronger or weaker" (p. 25). Again we see the principles of parsimony and replication in Fisher's words.

Although the presenters in this symposium are mainly educational researchers, it would seem that the issues discussed today impact all disciplines conducting empirical research and using statistical analyses. To date, thirteen journals that publish studies in education and psychology have adopted editorial policies requiring that effect sizes be reported (Thompson, in press) including:

Career Development Quarterly

Contemporary Educational Psychology

Educational and Psychological Measurement

Journal of Agricultural Education

Journal of Applied Psychology

Journal of Consulting and Clinical Psychology

Journal of Early Intervention

Journal of Experimental Education

Journal of Learning Disabilities

Language Learning

Measurement and Evaluation in Counseling and Development

Research in the Schools

The Professional Educator.

The conversation seems to be occurring in the sciences as well, evidenced by a recent article in the Journal of Wildlife Management (Anderson, Burnham, & Thompson, 2000).

What should be included in a primer on effect sizes? We reviewed not only papers on effect sizes but also papers on the most commonly reported statistical analyses in various journals. In their paper entitled "Twenty Years of Research Methods Employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research," Elmore and Woehlke (1998) found that the most frequent inferential statistical methods reported in the three journals combined between 1978 and 1997 were: Analysis of variance/covariance, multiple regression/correlation, and bivariate correlation. Analysis of variance/covariance, multiple regression, and bivariate correlation were the three most frequent methods used in the American Educational Research Journal from 1979 to 1983 (Goodwin & Goodwin, 1985b). Analysis of variance/covariance, bivariate correlation, t-test, and multiple regression were the four most frequent methods used in the Journal of Educational Psychology from 1979 to 1983 (Goodwin & Goodwin, 1985a). Kirk (1996) found that the three most frequently used inferential procedures in the 1995 volumes of the Journal of Applied Psychology; Journal of Educational Psychology; Journal of Experimental Psychology, Learning & Memory; and Journal of Personality and Social Psychology were analysis of variance, the t-test for means, and regression analysis. Given the frequency of the methods cited above, this primer will only discuss effect sizes relative to bivariate correlation, t-test, analysis of variance/covariance, and multiple regression/correlation.

Two Types of Effect Sizes

An agreed upon nomenclature regarding effect sizes or magnitude-of-effect statistics (Snyder & Lawson, 1993) might be useful. Rosenthal (1994) classified effect sizes into two families, the *r* family and the *d* family. The *r* family included the Pearson product moment correlation coefficient and the various squared indices of *r* and *r*-type quantities. The *d* family included mean difference and standardized mean difference indices. Maxwell and Delaney (1990) used the terms measures of association strength and measures of effect size for the *r* family and *d* family indices, respectively. Snyder and Lawson (1993) tried to further clarify the terms by stating that the measures of association strength involved proportions of variance ranging from 0 to 1 and the measures of effect size involved directly examining differences between means (p. 228).

Measures of Association Strength

The *r* Family of Effect Sizes

For studies using bivariate correlation, the Pearson product-moment correlation coefficient is used as the effect size estimate. For studies using multiple regression procedures, the coefficient of determination which is the obtained squared multiple correlation, R squared (R^2) is used. The coefficient of determination expresses the proportion of variance in the dependent variable accounted for by the linear combination of independent variables. In the first, second, and third editions of Multiple Regression in Behavioral Research Kerlinger and Pedhazur (1973) and Pedhazur (1982, 1997) have emphasized the use of the coefficient of determination in reporting regression results. Since the time of R. A. Fisher it has been known that analysis of variance/covariance and multiple regression are the same statistical techniques currently referred to as techniques based on the General Linear

Model. Eta squared (η^2) is the term used for analysis of variance/covariance analogous to the use of R^2 for multiple regression. Computationally R^2 and η^2 equal the sum of squares regression (between or explained or model) divided by the sum of squares total.

Pedhazur (1997) discussed the degree of overestimation of R^2 and the factors that affect the size of R^2: the ratio of the number of independent variables or predictors to the size of the sample and the value of R^2. Even though we cannot determine the exact amount of overestimation, it is possible to estimate the amount of "shrinkage" (p. 208) by applying the following formula:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{(N-1)}{(N-k-1)}$$

Snyder and Lawson (1993) refer to this adjustment as a "corrected" effect size estimate.

In addition to the concept of shrinkage relative to the population squared multiple correlation coefficient, Pedhazur (1997) discusses cross-validation when the purpose is "to determine how well a regression equation obtained in one sample performs in another sample from the same population" (p. 209). Again, we are concerned about the scientific principle of replication of findings.

Measures of Effect Size

The *d* Family of Effect Sizes

In 1969 Cohen proposed *d* which is the difference between population means divided by the average population standard deviation:

$$d = \frac{M_1 - M_2}{\sigma_{pooled}}$$

Glass (1976) defined the effect size as the difference between the experimental and control group means divided by the standard deviation of the control group:

$$\Delta = \frac{M_E - M_C}{S_C}$$

These two definitions of effect size are relevant when a t-test has been used and mean differences are tested. This type of effect size measure is particularly useful when the scale of measurement is meaningful to the researcher and reader. Cohen (1988) has provided guidelines for the interpretation of effect sizes: small *d*=. 2, medium *d*=. 5, and large *d*=. 8 (p. 24-27).

Other Measures of Effect Size

The purpose of this paper was to provide an introduction to effect size concepts. We refer you to in depth coverage of this topic in The Handbook of Research Synthesis (Cooper & Hedges, 1994), particularly Chapter 15, Combining Significance Levels by Betsy Jane Becker (1994); Chapter 16, Parametric Measures of Effect Size by Robert Rosenthal (1994); Chapter 17, Measures of Effect Size for Categorical Data by Joseph L. Fleiss (1994); Chapter 18, Combining Estimates of Effect Size by William R. Shadish and C. Keith Haddock (1994); Chapter 19, Fixed Effects Models by Larry V. Hedges (1994); and, Chapter 20, Random Effects Models by Stephen W. Raudenbush (1994). An additional excellent reference is Chapter 2, Statistical Methods in the Meta-analysis of Research on Gender Differences by Larry V. Hedges and Betsy Jane Becker in The Psychology of Gender: Advances through Meta-analysis edited by Janet Shibley Hyde and Marcia C. Linn (1986). Of course, the classic Statistical Methods for Meta-Analysis by

Larry V. Hedges and Ingram Olkin (1985) is a must read for a complete understanding of effect size issues in meta-analysis.

Recommendations

In 1969 John Tukey advocated the use of confidence intervals in psychology and the behavioral sciences. He said, "Amount, as well as direction is vital" (p. 86). Further, he noted "Measuring the right things on a communicable scale lets us stockpile information about amounts. Such information can be useful, whether or not the chosen scale is an interval scale" (p. 80).

In his seminal 1994 article entitled "The Earth Is Round ($p < .05$), Jacob Cohen made three recommendations: "First, don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist. Second, even before we, as psychologists, seek to generalize from our data, we must seek to understand and improve them. ...Thus, my third recommendation is that, as researchers, we routinely report effect sizes in the form of confidence limits" (p. 1001-1002). These recommendations seem as important and timely in 2001 as they did in 1994.

In an article published in 2001 Roger Kirk stated "I believe that science is best served when researchers focus on the size of effects and their practical significance. Questions regarding the size of effects are addressed with descriptive statistics and confidence intervals" (p. 214).

Is the idea of effect size a new concept? We think not. Even in 1951 a prominent statistician, Frank Yates, commented on the work of Fisher (1925/1951) as follows: "It has caused scientific research workers to pay undue attention to the results of the test of significance they perform on their data... and too little to the estimates of the magnitude of the effects they are estimating" (p. 32).

Conclusion

We want to end this presentation by returning to the scientific principles relevant to our discussion today: parsimony and replication. We would like to add a third principle, meaning or interpretability.

Should effect sizes be reported by researchers presenting findings of empirical studies in which statistical analyses were conducted? Yes.

What type of effect size should be reported? It depends. If the scale of measurement were meaningful to the discussion, then an effect size measure from the *d* family would seem appropriate. If the magnitude or strength of the effect were more meaningful, then an effect size measure from the *r* family would seem appropriate. The simple thing to remember is that there are formulas available to convert from one effect size measure to another regardless of type.

Researchers need to remember the warning of Rosnow and Rosenthal (1989) that "strength of effect is very context dependent. ...it is therefore important to recognize how the study characteristics might influence the size as well as one's interpretation of the magnitude-of-effect estimate" (p. 1280).

Possibly one of the most important issues to consider in educational and psychological research today is the quality of the measures we use. To enhance meaning and interpretability of research findings, it is essential that we study and report psychometric properties of variables and test scores for specific samples under varied conditions.

References

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. Journal of Wildlife Management, 64, 912-923.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 215-230). New York: Russell Sage Foundation.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Cooper, H., & Hedges, L. V. (Eds.). (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Elmore, P. B., & Woehlke, P. L. (1998). Twenty years of research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Fisher, R. A. (1951). Statistical methods for research workers. Edinburgh, Scotland: Oliver & Boyd. (Original work published 1925).

Fisher, R. A. (1960). Design of experiments (7th ed.). Edinburgh, Scotland: Oliver & Boyd.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 245-260). New York: Russell Sage Foundation.

Goodwin, L. D., & Goodwin, W. L. (1985a). An analysis of statistical techniques used in the Journal of Educational Psychology, 1979-1983. Educational Psychologist, 20 (1), 13-21.

Goodwin. L. D., & Goodwin, W. L. (1985b). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read educational research literature. Educational Researcher, 14 (2), 5-11.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 285-299). New York: Russell Sage Foundation.

Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds.), The psychology of gender: Advances through meta-analysis (pp. 14-50). Baltimore, MD: The Johns Hopkins University Press.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press, Inc.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement, 61, 213-218.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston, Inc.

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and analyzing data: A model comparison perspective. Belmont, CA: Wadsworth.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston, Inc.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 301-321). New York: Russell Sage Foundation.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 231-244). New York: Russell Sage Foundation.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 261-281). New York: Russell Sage Foundation.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Thompson, B. (in press). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? Journal of Counseling and Development.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24, 83-91.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46, 19-34.